EasyChair Preprint

№ 12258

Comparative Analysis of Housing Price Prediction of Dhaka City Using Machine Learning

Md Saiful Islam Sajol and Mohammad Nizam Uddin

February 29, 2024

Comparative Analysis of Housing Price Prediction of Dhaka City using Machine Learning

Md Saiful Islam Sajol *, Mohammad Nizam Uddin†
*Louisiana State University, Louisiana, USA,
†Bangladesh University of Engineering and Technology, Dhaka, Bangladesh
Email: msajol1@lsu.edu, nizamuddin.ce.buet@gmail.com

***Abstract*—The housing market is rapidly expanding, making it crucial to forecast housing prices for both company and individual consumers. However, several factors affect house price variations. As Bangladesh is an overpopulated country, the sale price of real estate is influenced by several interrelated factors. The size, location, and amenities of the property are important variables that could determine the price. We examined about 8000 houses in the Dhaka and Chittagong Region in Bangladesh as a case study and discussed how the increase in housing prices could vary by each of the contributing components. In this research, we conduct an extensive analysis and investigation of twelve machine learning methods for housing price prediction. Our study offers a comprehensive study for assessing the effectiveness and reliability of machine learning models for predicting home prices. The results showed LGBM to be the second best model with an R2 equal to 85% and XGB to be the best model with an R2 equal to 94%.**

***Index Terms*—Machine Learning, Housing, Price Prediction, KNN, Random Forest**

I. Introduction

The world of real estate is a complex and often contentious realm, where the stakes are high, and buyers are left pondering the ever-fluctuating housing prices. As economies expand and urbanization gathers pace, the real estate market is witnessing an unprecedented boom. In this landscape, the price tags attached to homes become a topic of utmost concern, as location, amenities, and neighborhood allure all play their part in shaping the value. Understanding the trajectory of housing prices is paramount for analyzing the real estate market and making informed decisions based on specific locations and desirable amenities, such as parking and neighborhood features. While extensive research has been dedicated to housing price modeling, there is an ongoing need for comprehensive comparative studies to accurately assess fluctuations in house prices [1]. Predicting these prices proves to be an arduous task, given the dynamic nature of contributing factors and their vulnerability to regulatory influences. In this pursuit, the ability to gain future insights into housing market trends not only cultivates trust among potential investors but also empowers individuals to engage in realistic financial planning and secure their future. Accurate predictions regarding the housing market play a vital role in facilitating informed decision-making, reducing uncertainties, and maximizing investment opportunities [2]. Machine learning algorithms have been very useful in processing multi-modal data [11] and optimization [13], [14].

Despite significant research dedicated to housing price modeling, there remains a need for further comparative studies to assess house price fluctuations accurately. Modeling makes use of machine learning methods, which enable computers to learn from data and predict new data. Machine learning entails the provision of valid datasets, and predictions are then based on them. The machine itself learns the potential significance of a given event for the system as a whole based on the data it has already loaded, and it predicts the outcome appropriately. The list of contemporary uses for this technique is vast and includes forecasting stock prices, the likelihood of an earthquake, company sales, and many more.

In this work, we have attempted to compare different house price prediction machine learning models using real estate data from Bangladesh which is publicly available. The Support Vector Machine, Decision Tree, KNN, LGBM, XGB, Linear Regression, Random Forest, Extra Trees Regressor, Bayesian Ridge, Kernel Ridge, and Elastic Net were the twelve prediction models we took into consideration. The use of evaluation measures has also been used in comparative research. Once the model and the data are well-fit, we use them to predict the financial value of that specific Bangladeshi housing property.

II. Related Works

The economic progress and quality of life of the country depend greatly on accurate house price predictions. In the past, the grey model (GM) was commonly used for predicting home prices, but it had limitations such as requiring a small number of samples and showing only a monotonous increase or decrease. To overcome these limitations, authors [3], propose the use of Support Vector Machine (SVM), which avoids overfitting and reduces prediction errors by following the principle of structural risk minimization. There are various methods available for optimizing SVM parameters, with grid algorithms and genetic algorithms being the most commonly used. It suggests that using a genetic algorithm can optimize SVM parameters simultaneously and in less time. It also introduces the G-SVM algorithm as a reliable and efficient method for predicting housing prices, addressing the limitations of previous approaches.

A kernel function, kernel parameters, a soft margin constant, and an insensitive loss parameter must all be selected to produce the best SVM forecasting model. Ant colony algorithms, grid algorithms, and genetic algorithms are the most widely

used optimization approaches. However, the grid algorithm method is computationally intensive, time-consuming, and has low learning accuracy; the ant colony algorithm method has disadvantages such as initial pheromone scarcity, long-term searching, and local best solution; the genetic algorithm is operation complex and different issues need to design different crossover or mutation. In this research [4], the author introduced particle swarm optimization (PSO) to have the extensive capability of global optimization for its simple concept, easy implementation, and fast convergence. To evaluate the predicting capabilities of the PSO-SVM model and the BP neural network, Chongqing real estate sample data are used. According to the experimental findings, PSO-SVM has a higher forecasting accuracy than the BP neural network for real estate price forecasting.

In this work [5], PCA is used as a data transformation approach to extract significant components, which are then combined with SVM and various regression models to identify the most relevant ones. First, the raw data is cleaned up and packaged into a dataset that is suitable for analysis. Then, using Stepwise and PCA approaches, data reduction and transformation are conducted. To find the best solution, many techniques are then used and assessed. These include Linear Regression, Polynomial Regression, Regression Tree, Neural Network, and SVM. The evaluation phase indicates that the combination of Step-wise and the SVM model is a competitive approach.

In [6], forecasting the co-movements in housing prices is generated by controlling for a wide range of predictors, such as factors generated from a large macroeconomic data set, oil shocks, and financial market-related uncertainties. Bayesian dynamic factor model and random forest machine learning approach are used to account for various predictors and non-linearities.

This study [7], examined about 500 houses in the Boston area and discussed how the growth in property prices could differ depending on each of the contributing factors. A comparison of the accuracy of all the models is conducted using different machine learning (ML) regressors in the dataset. However, scarcity of data has led to poor model performance in some areas. The results showed that the voting regressor was rated second highest, with an R2 of 0.87, while the random forest was rated best, with an R2 of 0.88. According to the findings of multivariate exploratory data analysis, the average number of rooms and proportion of the population with lower socioeconomic class had the most effects on price range estimates.

In this research [8], as the primary location Mumbai is considered and real-time house prices for various localities in and around Mumbai are predicted. Linear regression, forest regression, and boosted regression are all used to their full potential by the system. With the usage of neural networks, the algorithm's efficiency has been significantly boosted.

In [9], According to data gathered in 2016, Bengaluru is quickly expanding both in terms of area and population though The data set's attributes are insufficient to adequately reflect Bengaluru house prices. Based on the elements that influence pricing, an effort has been made to build a prediction model for assessing the price. Lasso and Ridge regression models, support vector regression, multiple linear regression (Least Squares), and boosting algorithms like Extreme Gradient Boost Regression (XG Boost) are some of the regression techniques used in modeling explorations.

In this study [10], a dataset is built with information on Taiwan's housing features and macroeconomic conditions. This dataset is used to test different strategies for predicting housing prices based on the deep learning algorithms BPNN and CNN. The best model for the suggested methods is regarded to be CNN, where R2 is greater. CNN is typically used in the field of image processing, but this study shows that it can also be effective in time series prediction. Because of this, this research demonstrates that convolutional neural networks are appropriate for regression prediction of home prices and can efficiently learn time-series data.

III. Methodology

We collected the dataset from Kaggle house pricing data which includes a comprehensive collection of house listings from different Bangladeshi cities and areas (Dhaka, Chittagong), information on location, property type, size, amenities, and price data spanning from June 3, 2020, to January 23, 2023 [12]. The dataset is regularly updated, and carefully segmented, focusing on specific target values and giving a precise and current picture of Bangladesh's housing sector. Our goal was to correctly forecast house pricing.

A. Dataset Collection

Our findings from this dataset provide valuable insights into the relationships between different house pricing variables and offer valuable information for forecasting house pricing and understanding house pricing patterns in the Dhaka and Chittagong regions. There are 11 feature variables and 8000 samples in the dataset. With the help of the provided features, the objective is to estimate the value of the housing price. The extracted features are given below:

- Title: Name of the house property
- beds: Number of bedrooms
- baths: Number of bathrooms
- Area: Area of the house property in square feet
- Address: Housing property's location
- Purpose: Whether the house property is for living or rent
- Floor Plan: Image of the architectural design of the floor
- URL: URL of the house listing on the website
- Last Update: Last update date of the house property
- Price: Price in Bangladeshi Taka

B. Data Pre-processing

Pre-processing was done on the dataset to deal with missing values, outliers, and categorical variables. Missing values are filled using forward fill, which has replaced the NULL values with the value from the previous row. Some columns (Floorplan, URL) won't be required when conducting data analysis. In this regard, these two columns have been removed.

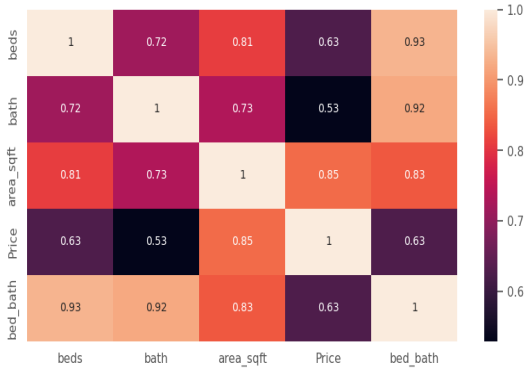

Fig. 1. Correlation heatmap between the features

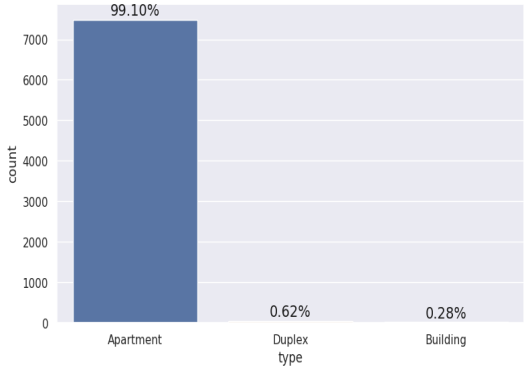

Fig. 2. House Type in Dhaka City

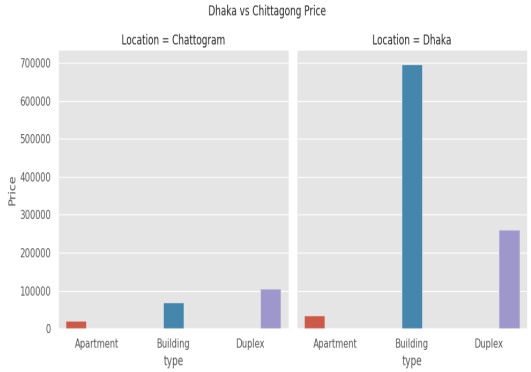

Fig. 3. Price difference between Dhaka and Chittagong

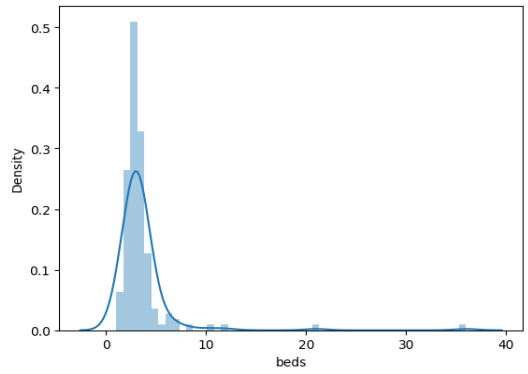

Fig. 4. Mean Price Density for different beds

C. Data Visualisation and Analysis

Data visualization is the graphic or pictorial display of information. From the fig-1 heatmap, we have found the correlation that tells how much variable changes concerning the other. From fig-2, We can observe that in Dhaka, the vast majority of homes are apartments, with a small minority being duplexes and buildings. From Fig -3, The average price differential between Chittagong and Dhaka is fairly large. Compared to Chittagong, Dhaka is significantly more expensive. From Figure 4, it contains a new Series object that represents the mean price for each unique value in the 'beds' column. The index of the Series will correspond to the unique values in the 'beds' column, and the values will represent the mean prices associated with each group. Similarly, figure -5 contains the mean price density along with bath numbers.

D. Model Training

Following data pre-processing and data visualization, the entire dataset is divided into two sections: the train set, which is used to train the model, and the test set, which is used to test the model. Various machine learning methods can be used to predict the house price. In our work, we propose twelve models: Support Vector Machine, Decision Tree, K-Nearest Neighbor, LGBM, XGBoost, Linear Regression, Random Forest Regression, Extra Trees Regressor, Bayesian Ridge, Kernel Ridge, SGD Regressor, and Elastic Net to see which one gives better performance.

IV. Experiment And Analysis

We evaluated the model's performance using metrics: R-squared, Root Mean Square Error, MAE, and MSE. R-squared (also known as the coefficient of determination) is a quadratic statistical criterion that measures how well the real target data match the fitted regression line. The variance between the anticipated target variable and the actual rent price is depicted

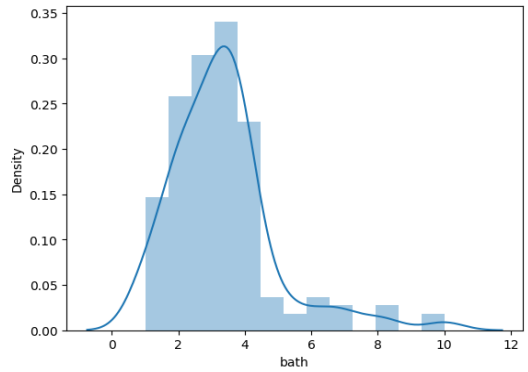

Fig. 5. Mean Price Density for different baths

TABLE I
MODEL COMPARISON AND EVALUATION RESULT

Algorithm	R-Squared	RMSE	MAE	MSE
Support Vector Machine	5%	0.60	0.19	0.368
Decision Tree	57%	0.387	0.985	0.149
K- Nearest Neighbor	74%	0.2988	0.9231	0.893
LGBM	85%	0.224	0.822	0.50
XGB	94%	0.144	0.66	0.207
Linear Regression	69%	0.32	0.13	0.10
Extra Trees Regressor	68%	0.421	0.1313	0.177
Bayesian Ridge	76.5%	0.364	0.146	0.132
Kernel Ridge	77%	0.360	0.144	0.129
Elastic Net	74.5%	0.380	0.161	0.144
Random Forest	74%	0.299	0.93	0.89

in the paper using R-squared. Therefore, the smaller the MAE and the higher the R-squared, the better our model fits the data. In addition, Lower values of MSE, and RMSE indicate better fit.

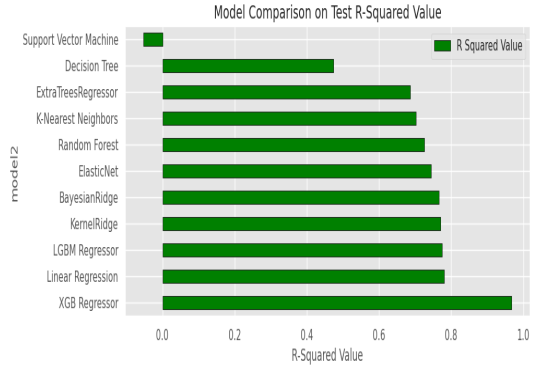

Fig. 6. R-Squared Value of the models

In comparison to all other algorithms for predicting property prices, our study shows that the XG Boost model has a high accuracy value. Bangladesh's Real Estate Market Dataset, which was compiled from publicly available datasets, is used to calculate the accurate value of the algorithm using the Root Mean Square Error and [RMSE] and R2-squared value primarily. For the proposed methods, KNN, LGBM, and Random Forest showed satisfactory performance. But XG Boost should be carried out to predict house price as its R-squared value is 94% and RMSE value is 0.144.

From the table, we can easily perform the comparison of different algorithms clearly to find the best among them where XGB outperforms all other models whose R-squared value is 94%. XGB models' RMSE value is 0.144 which is less compared to all other respective algorithms. In Figure 5

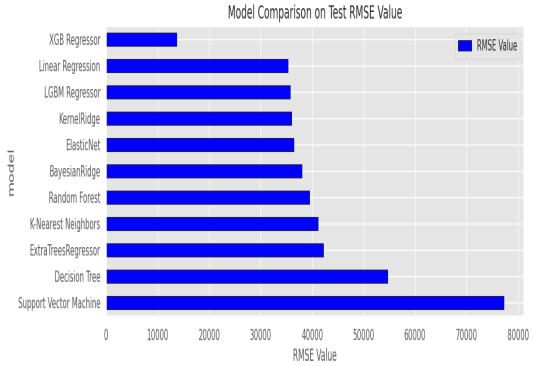

Fig. 7. RMSE Value of the models

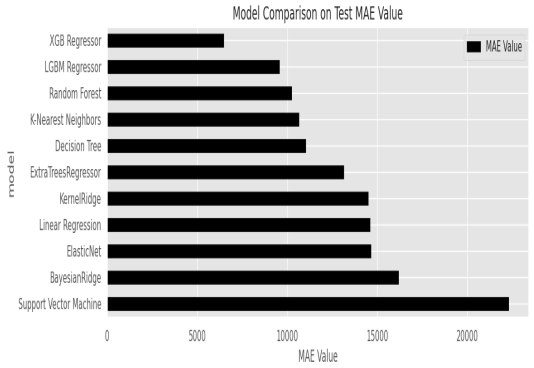

Fig. 8. MAE Value of the models

and Figure 6, the graphical representation of all the different regression techniques listed represented the accuracy of the models.

V. FUTURE DIRECTION

In the future, this research can be expanded to enhance housing price prediction models in Bangladesh. First, integrating additional features like proximity to amenities, crime rates, and socioeconomic indicators could improve model accuracy. Second, exploring advanced modeling techniques such as deep learning algorithms and ensemble methods could further enhance predictive performance. Additionally, incorporating temporal analysis and forecasting would capture market trends. Future studies should focus on data quality and preprocessing techniques, and validating models with real-world housing transaction data would refine their applicability. Pursuing these directions can contribute to more accurate and robust housing price prediction models, aiding stakeholders in informed decisions.

VI. CONCLUSION

In conclusion, this study aimed to compare various machine learning models for house price prediction using publicly available real estate data from Bangladesh. The Support Vector Machine, Decision Tree, KNN, LGBM, XGB, Linear Regression, Random Forest, Extra Trees Regressor, Bayesian

Ridge, Kernel Ridge, SGD Regressor, and Elastic Net models were evaluated and analyzed. Through the use of evaluation measures, we assessed the performance of these models and identified their strengths and weaknesses. By carefully fitting the models to the dataset, we obtained well-fitting models that could effectively predict the financial value of Bangladeshi housing properties. This research provides valuable insights into the potential of machine learning algorithms in the context of real estate and housing price prediction in Bangladesh. Overall, this work contributes to the growing field of real estate analytics and provides a foundation for future studies aiming to develop more accurate and robust models for housing price prediction in Bangladesh. The application of machine learning techniques in the real estate sector holds great promise for assisting stakeholders, such as buyers, sellers, and investors, in making informed decisions related to property valuation and investment strategies

REFERENCES

[1] https://www.bankrate.com/real-estate/housing-market-5-year-forecast/

[2] https://www.idealhome.co.uk/news/good-move-housing-market-10-year-forecast-265707

[3] Gu J, Zhu M, Jiang L. Housing price forecasting based on genetic algorithm and support vector machine. Expert Systems with Applications. 2011 Apr 1;38(4):3383-6

[4] Xibin Wang, Junhao Wen, Yihao Zhang, Yubiao Wang, Real estate price forecasting based on SVM optimized by PSO, Optik,Volume 125, Issue 3,2014,Pages 1439-1443,ISSN 0030-4026

[5] T. D. Phan, "Housing Price Prediction Using Machine Learning Algorithms: The Case of Melbourne City, Australia," 2018 International Conference on Machine Learning and Data Engineering (iCMLDE), Sydney, NSW, Australia, 2018, pp. 35-42, doi: 10.1109/iCMLDE.2018.00017.

[6] Gupta, Rangan, et al. "Machine learning predictions of housing market synchronization across US states: the role of uncertainty." The Journal of Real Estate Finance and Economics (2022): 1-23.

[7] Khosravi, M.; Arif, S.B.; Ghaseminejad, A.; Tohidi, H.; Shabanian, H. Performance Evaluation of Machine Learning Regressors for Estimating Real Estate House Prices. Preprints.org 2022, 2022090341. https://doi.org/10.20944/preprints202209.0341.v1

[8] A. Varma, A. Sarma, S. Doshi and R. Nair, "House Price Prediction Using Machine Learning and Neural Networks," 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT), Coimbatore, India, 2018, pp. 1936-1939, doi: 10.1109/ICICCT.2018.8473231.

[9] Manasa, J., Radha Gupta, and N. S. Narahari. "Machine learning based predicting house prices using regression techniques." 2020 2nd International conference on innovative mechanisms for industry applications (ICIMIA). IEEE, 2020.

[10] Zhan C, Wu Z, Liu Y, Xie Z, Chen W. Housing prices prediction with deep learning: an application for the real estate market in Taiwan. In2020 IEEE 18th International Conference on Industrial Informatics (INDIN) 2020 Jul 20 (Vol. 1, pp. 719-724). IEEE.

[11] M. S. Islam, M. S. Rahman, and M. A. Amin, "Beat Based Realistic Dance Video Generation using Deep Learning," 2019 IEEE International Conference on Robotics, Automation, Artificial-Intelligence and Internet-of Things (RAAICON), Dhaka, Bangladesh, 2019, pp 43-47. doi:10.1109/RAAICON48939.2019.22

[12] https://www.kaggle.com/datasets/ijajdatanerd/property-listing-data-in-bangladesh

[13] A. S. M Jahid Hasan, M. S. Rahman, M. S. Islam and J. Yusuf, "Data Driven Energy Theft Localization in a Distribution Network," 2023 International Conference on Information and Communication Technology for Sustainable Development (ICICT4SD), Dhaka, Bangladesh, 2023, pp. 388-392, doi: 10.1109/ICICT4SD59951.2023.10303520.

[14] A. S. M. Jahid Hasan, J. Yusuf, M. S. Rahman and M. S. Islam, "Electricity Cost Optimization for Large Loads through Energy Storage and Renewable Energy," 2023 International Conference on Information and Communication Technology for Sustainable Development (ICICT4SD), Dhaka, Bangladesh, 2023, pp. 46-50, doi: 10.1109/ICICT4SD59951.2023.10303409.